DOCUMENT RESUME

ED 473 524 TM 034 732

AUTHOR	Fox, Jean-Paul
TITLE	Multilevel IRT Using Dichotomous and Polytomous Response Data. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO	RR-02-11
PUB DATE	2002-00-00
NOTE	39p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. E-mail: Fox@edte.utwente.nl.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Bayesian Statistics; Estimation (Mathematics); *Item Response Theory; Markov Processes; Monte Carlo Methods; Simulation
IDENTIFIERS	*Dichotomous Responses; Multilevel Analysis; *Polytomous Items

ABSTRACT

A structural multilevel model is presented in which some of the variables cannot be observed directly but are measured using tests or questionnaires. Observed dichotomous or ordinal politicos response data serve to measure the latent variables using an item response theory model. The latent variables can be defined at any level of the multilevel model. A Bayesian procedure, the Markov Chain Monte Carlo (MCMC) procedure, to estimate all parameters simultaneously is presented. It is shown that certain model checks and model comparisons can be done using the MCMC output. The techniques are illustrated using a simulation study, and applications involving a student's achievements on a mathematics test and test results regarding management characteristics of teachers and principals. (Contains 2 figures, 4 tables, and 45 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM TM 0072

ED 473 524

Multilevel IRT Using Dichotomous and Polytomous Response Data

Research Report 02-11

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Nelissen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Jean-Paul Fox

TM034732

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

BEST COPY AVAILABLE

Multilevel IRT Using Dichotomous and Polytomous Response Data

Jean Paul Fox

° Requests for reprints should be sent to Jean-Paul Fox, Department of Educational Measurements and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, THE NETHERLANDS. Email: Fox@edte.utwente.nl

Abstract

A structural multilevel model is presented where some of the variables cannot be observed directly but are measured using tests or questionnaires. Observed dichotomous or ordinal polytomous response data serve to measure the latent variables using an item response theory model. The latent variables can be defined at any level of the multilevel model. A Bayesian procedure, MCMC, to estimate simultaneously all parameters is presented. It is shown that certain model checks and model comparisons can be done using the MCMC output. The techniques are illustrated using a simulation study and an application involving student's achievements on a mathematic test and test results regarding management characteristics of teachers and principles.

Key words: Gibbs sampler, graded response model, hierarchical linear models, item response theory, Markov Chain Monte Carlo, measurement error, Metropolis-Hastings, multilevel IRT, multilevel model, two-parameter normal ogive model.

Introduction

School effectiveness research is a major topic in education, especially in light of the concern for evaluation of differences in achievement and accountability. Main interest is put in identifying the characteristics of effective schools and criteria for measuring effectiveness. The methods of measuring school effectiveness have been changed radically with the development of multilevel analysis. The hierarchical structure of educational systems emphasizes the necessity of multilevel modeling. Multilevel analysis enables that the data are treated in an appropriate manner, instead of being reduced to a single level. The differences between classes and schools can be taken into account properly, rather than aggregated arbitrarily. In this framework, most of the variance is explained by student background variables, such as intelligence and socio-economic status, other parts of the variance can be explained by class or school factors. Applications of multilevel models to educational data can, for example, be found in Bock (1989) and Goldstein (1995).

In a standard application in school effectiveness research there are several schools, with varying numbers of students, and each student has a test score. Interest is focused on the effect of student and school characteristics on the students' achievements. A major component in the analysis is the use of achievement scores as a measure of effectiveness. Most often, schools are compared in terms of the achievements of the pupils, and test scores are used to represent these achievements. Students' achievements cannot be observed directly but are observed by manifest variables or proxies. It may also be possible that some explanatory variables on different levels are observed by manifest variables, such as, intelligence, socio-economic status, or community loyalty. Obviously, errors of measurement are inherent to manifest variables. Traditionally, the manifest variables are used in further analyses as fixed and known entities. An important deficiency is that the measurement error associated with the test scores are ignored. This error can have an effect on the estimates of the parameters of the multilevel model, that is, the standard errors of the parameters are underestimated. In general, the use of unreliable variables leads to biased estimation of the regression coefficients and the resulting statistical inference can be very misleading.

This problem can be handled by extending an item response theory (IRT) model to a multilevel item response theory model consisting of a latent variable assumed to be the outcome in a regression analysis. This model has already become an attractive alternative to the traditional multilevel models. This model is often presented as two or three-level formulation of an item response model, that is, a multilevel regression model is imposed on the ability

parameter in an item response model. Verhelst and Eggen (1989) and Zwinderman (1991, 1997) defined a structural model for the one parameter logistic model and the Rasch model with observed covariates assuming the item parameters are known. Zwinderman also illustrates the possibility for modeling differential item functioning. Adams, Wilson and Wu (1997) and Raudenbush & Sampson (1999) discussed a two and three level hierarchical logistic regression model which can be seen as a Rasch model embedded within a hierarchical structure. The first level of the multilevel model describes the relation between the observed item scores and the ability parameters. This two and three-level model can be estimated in HLM 5 (Raudenbush, Bryk, Cheong, & Congdon, 2000). Kamata (2001) defined the multilevel formulation of the Rasch model as a hierarchical generalized linear model that can be estimated within the HLM software. Also, Maier (2001) defines a Rasch model with a hierarchical model imposed on the person parameters but without additional covariates. Fox and Glas (2001, 2002) extended the two-parameter normal ogive model by imposing a multilevel model, with covariates on both levels, on the ability parmeters. This multilevel IRT model describes the link between dichotomous response data and a latent dependent variable within a structural multilevel model. They also showed how to model latent explanatory variables within a structural multilevel model using dichotomous response data.

All these developed models can handle dichotomous response data, that is, the Rasch model or the normal ogive model is used as an item response model for measuring the latent variables. But data collected from respondents utilizing questionnaires and surveys are often polytomous. For example, the use of Likert items on questionnaires is frequently used in educational and psychological measurement. Treating the polytomous data as continuous and ignoring the ordinal discrete nature of the data can lead to incorrect conclusions (Lee, Poon, & Bentler, 1992). On the other hand, transforming the polytomous data to dichotomous data, by collapsing response categories to enforce dichotomous outcomes, leads to a loss of information contained in the data. The best way is to extend the models to handle polytomous data measuring one latent ability. In the present paper, attention is focused on measuring latent dependent and independent variables of a multilevel model where manifest variables, consisting of binary, ordinal, or graded reponses, are available. This extension makes it possible to model relationships between observed and latent variables on different levels using dichotomous and polytomous item response theory models to describe the relationship between the test performances and the latent variables. That is, relationships between abilities of students underlying the test and other observed variables or other measurements of some individual or group characteristics can be analyzed taking into account the errors of measurement using

dichotomous or polytomous indicators.

It will be shown that adopting a fully Bayesian framework results in a straightforward and easily implemented estimation procedure. That is, a Markov Chain Monte Carlo method will be used to estimate the parameters of interest. Computing the posterior distributions of the parameters involves high dimensional integrals but these can be carried out by Gibbs sampling (Gelfand, Hills, Racine-Poon, & Smith, 1990, Gelman, Carlin, Stern, & Rubin, 1995). Within this Bayesian approach, all parameters are estimated simultaneously and goodness-of-fit statistics for evaluating the posited model are obtained.

After this introduction, the model will be presented. In the next section, prior choices and the estimation procedure will be discussed. Then, several criteria, as the posterior predictive check, pseudo-Bayes factor and the marginal likelihood, are introduced to assess the model fit. In the following section a simulation study and a real data example will be given. The last section contains a discussion and suggestions for further research.

Model Description

Educational or psychological tests are used for measuring variables as intelligence and arithmetic ability which cannot be observed directly. Interest is focused on the knowledge or characteristics of students given some background variables but only the performance on a set of items is recorded. Item response theory models can be used to describe the relationship between the abilities and the responses of the examinees to the items of the test to assess the abilities of the examinees. The class of item response theory (IRT) models, is based on the charateristics of the items in the test. The dependence of the observed responses to binary or polytomously scored items on the latent ability is specified by item characteristic functions. In case of binary items, the item characteristic function is the regression of item score on the latent ability. Under certain assumptions it is possible to make inferences about the latent ability from the observed item responses using the item response functions. In specific, the probability of a student corresponding correct to an item k $(k = 1, \ldots, K)$, is given by

$$P(Y_k = 1 \mid \theta, a_k, b_k) = \Phi(a_k\theta - b_k), \tag{1}$$

where $\Phi(.)$ denotes the standard normal cumulative distribution function, and a_k and b_k are the discrimination and difficulty parameter of item k. Below, the parameters of item k will also be denoted by $\boldsymbol{\xi}_k$, $\boldsymbol{\xi}_k = (a_k, b_k)^t$. The relation between the underlying latent ability, θ, and the

dichotomous outcomes can also be explained as follows. Assume a latent independent random variable Z_k normally distributed with mean $a_k\theta - b_k$ and variance 1. Further, the response Y_k is the indicator of Z_k being positive. Thus, a correct response on item k is obtained if a positive value is drawn from this normal distribution with mean $a_k\theta - b_k$ and variance 1. In Appendix A it will be shown that the introduction of the latent random variables simplifies the implementation of the MCMC algorithm.

The transition to polytomous scored items can be done by defining the polytomous response, Y, as an indicator of Z falling into one of the response categories. Or the other way around, classifying the latent variable Z into more than two categories is done by the cutoff or threshold parameters κ. In this case, the latent variable Z is defined as

$$Z_k = a_k\theta + \epsilon_k \tag{2}$$

where ϵ_k is assumed to be standard normal distributed. When the value of the latent variable Z_k falls between the thresholds κ_{kc-1} and κ_{kc}, the observed response on item k is classified into category c. The ordering of the response categories is displayed as follows,

$$-\infty < \kappa_{k1} \leq \kappa_{k2} \leq \ldots \leq \kappa_{kC_k}, \tag{3}$$

where there are C_k categories. Notice that the number of categories may differ per item. Here, for notational convenience, $\kappa_0 = -\infty$ and the upper cutoff parameter $\kappa_{kC_k} = \infty$ for every item k $(k = 1, \ldots, K)$. The probability that an individual, given some underlying latent ability, θ, obtains a grade c, or gives a response falling into category c on item k is defined by

$$P\left(Y_k = c \mid \theta, a_k, \kappa_k\right) = \Phi\left(\kappa_{kc} - a_k\theta\right) - \Phi\left(\kappa_{kc-1} - a_k\theta\right), \tag{4}$$

where $\Phi(.)$ denotes the standard normal cumulative distribution function. This item response model, called the graded response model or the ordinal probit model, for polytomous scored items have been used by several researchers, among others, Johnson and Albert (1999), Muraki and Carlson (1995) and, Samejima (1969). Notice, the cumulative probability model for ordinal polytomous response data, formula (4), implies that the slope parameters of different categories within an item must be constrained to be equal, see Mellenbergh (1995).

The measurement model is sometimes of interest in its own right, but here attention is focused on relations between latent variables and other observed variables. The structural

multilevel model defines the relations between the underlying latent variables and other important variables at different levels. In the present paper, a sample of clusters, say schools, indexed $j = 1, \ldots, J$ is considered. A total of N Individuals, labeled $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, are nested within clusters. Consider at Level 1, an observed or latent dependent variable ω and Q covariates, where $Q - q$ covariates are observed without an error, $\mathbf{X}_{ij}$, and q latent covariates $\boldsymbol{\theta}_{ij}$. At Level 2, S covariates are considered, containing $S - s$ covariates observed without an error, $\mathbf{W}_j$ of dimension $(Q \times (S - s))$, and s latent covariates $\boldsymbol{\zeta}_j$ of dimension $(Q \times s)$. This corresponds with the following structural multilevel model

$$\begin{aligned} \omega_{ij} &= \boldsymbol{\beta}_j' \left[\mathbf{X}_{ij}, \boldsymbol{\theta}_{ij}\right] + e_{ij} \\ \boldsymbol{\beta}_j &= \boldsymbol{\gamma}' \left[\mathbf{W}_j, \boldsymbol{\zeta}_j\right] + \mathbf{u}_j \end{aligned} \tag{5}$$

where $e_{ij} \sim N(0, \sigma^2)$, and $\mathbf{u}_j \sim N(0, \mathbf{T})$. Notice, the coefficients, regarding the observed and latent covariates at Level 1, vary over Level 2 clusters and are both regressed on observed covariates $\mathbf{W}$ and latent covariates $\boldsymbol{\zeta}$.

Both measurement models, the normal ogive and the graded response model are not identified. The models are overparameterized and require some restrictions on the parameters. The most common way is to fix the scale of the latent ability to a standard normal distribution. As a result, the multilevel IRT model, (5), is identified by fixing the scale of the latent abilities. Another possibility is to impose identifying restrictions on the item parameters. In case of the of the normal ogive model, this can be done by imposing the restriction, $\prod_k \alpha_k = 1$, and , $\sum_k b_k = 0$.

Besides the regression among latent variables, it is possible to incorporate latent variables at the lower level as a predictor of latent abilities at the higher level. In Fox and Glas (2002), an example is given of a covariate representing adaptive instruction of teachers, measured with a test consisting of 23 dichotomous items, predicting the abilities of the students. Below, an example will be given of school climate reflecting students' math abilities, where school climate will be measured with 23 polytomous items and the math abilities by 50 dichotomous items.

Handling response error in both the dependent and independent variables in a multilevel model using item response theory has some advantages. Measurement error can be defined locally as the posterior variance of the ability parameter given a response pattern resulting in a more realistic, heteroscedastic treatment of the measurement error. Besides the fact that in IRT reliability can be defined conditionally on the value of the latent variable offers

the possibility of separating the influence of item difficulty and ability level, which supports the use of incomplete test administration designs, optimal test assembly, computer adaptive testing and test equating. Further, it is possible handle various kinds of item responses to assess the ability of interest without simplifying assumptions regarding the discrete nature of the responses.

Parameter Estimation

Let $\mathbf{y}$ be the matrix of observed data, where $\mathbf{y} = (\mathbf{y}^{\omega}, \mathbf{y}^{\theta}, \mathbf{y}^{\zeta})$ denote the observed data in measuring the latent abilities ω, θ and ζ, respectively. The likelihood of the parameters of interest of model (5) is a product of the likelihood for the J groups, that is,

$$
\begin{aligned}
l(\boldsymbol{\xi}, \sigma^2, \gamma, \mathbf{T} \mid \mathbf{y}) = \prod_j \int \Bigg[\prod_{i|j} \int f(\mathbf{y}_{ij}^{\omega} \mid \boldsymbol{\xi}^{\omega}, \omega_{ij}) \, p(\omega_{ij} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\beta}_j, \sigma^2) \\
\prod_q \left[\int g_q(\mathbf{y}_{qij}^{\theta} \mid \boldsymbol{\xi}_q^{\theta}, \theta_{qij}) \, p\left(\theta_{qij}; \mu_{\theta_q}, \sigma_{\theta_q}^2\right) d\theta_{qij} \right] d\omega_{ij} \Bigg] \\
p(\boldsymbol{\beta}_j \mid \zeta_j, \gamma, \mathbf{T}) \prod_s \left[\int h_s\left(\mathbf{y}_{sj}^{\zeta} \mid \boldsymbol{\xi}_s^{\zeta}, \zeta_{sj}\right) p\left(\zeta_{sj}; \mu_{\zeta_s}, \sigma_{\zeta_s}^2\right) d\zeta_{sj} \right] d\boldsymbol{\beta}_j, \qquad (6)
\end{aligned}
$$

where $f(\mathbf{y}_{ij}^{\omega} \mid \boldsymbol{\xi}^{\omega}, \omega_{ij})$ is an IRT model, specifying the probability of the observing response pattern $\mathbf{y}_{ij}^{\omega}$ as a function of the ability parameter ω_{ij} and item parameters $\boldsymbol{\xi}^{\omega}$. Further, $g_q(\mathbf{y}_{qij}^{\theta} \mid \boldsymbol{\xi}_q^{\theta}, \theta_{qij})$ is an IRT model for q^{th} latent explanatory variable on Level 1, θ_{qij}, using dichotomous or polytomous response data $\mathbf{y}_{qij}^{\theta}$ and item parameters $\boldsymbol{\xi}_q^{\theta}$. In the same way, $h_s\left(\mathbf{y}_{sj}^{\zeta} \mid \boldsymbol{\xi}_s^{\zeta}, \zeta_{sj}\right)$ is an IRT model for the s^{th} latent explanatory variable on Level 2, ζ_{sj}, using the observed data $\mathbf{y}_{sij}^{\zeta}$ and item parameters $\boldsymbol{\xi}_s^{\zeta}$. Here, it is assumed that the latent explanatory variables θ and ζ are both mutually independent. It is possible to model correlated latent covariates at the same level. Fox and Glas (2002) transformed the parametrization of the latent variables in such a way that the latent variables are independent. The same procedure can be applied.

Computing expectations of marginal distributions using, for example, Gauss-Hermite quadrature is difficult and becomes infeasible when the number of latent variables is increasing. Furthermore, the frequentist methods that rely on large-sample theory will not be appropriate when the sample size regarding the number of items, the number of respondents or the number of group sizes is small. In specific, the asymptotic properties are easily violated in case of small groups of discrete responses. On the other hand, a Bayesian approach has

the advantage that computations for estimation can be based on MCMC methods, which circumvent the computation of high dimensional integrals. Moreover, the Bayesian approach gives the possibility to model all dependencies among variables and all sources of uncertainty.

Priors

Bayesian procedures require the specification of priors, that is, in order to form a posterior density, all prior distributions of all model parameters must be specified. Diffuse proper priors will be used to reflect vague beliefs about the parameter values. In formula (6), it is assumed that the latent abilities are drawn from a normal distribution. As mentioned, identification of the model can be done by specifying the scale of the latent variables, for example, by stating that each latent variable is standard normal distributed. The Bayesian approach has the advantage that the identification of the model can be done by defining an appropriate prior for the latent abilities.

The normal ogive model has two item-specific parameters, a discrimination and a difficulty parameter, formula (1). The prior for the difficulty and discrimination parameter insured that each item had a positive discrimination index, and assumed independence between the item difficulty and discrimination parameter,

$$p(\boldsymbol{\xi}) = p(\mathbf{a})\, p(\mathbf{b}) \propto \prod_{k=1}^{K} I(a_k > 0)\, I(a_k, b_k \in A), \tag{7}$$

where A is a sufficiently large bounded interval. The prior for the item-parameters in the graded response model, formula (4), can be specified in the same manner. That is,

$$p(\boldsymbol{\xi}) = p(\mathbf{a})\, p(\boldsymbol{\kappa}) \propto \prod_{k=1}^{K} I(a_k > 0)\, I(a_k, \kappa_{k1}, \ldots, \kappa_{kC_k} \in A), \tag{8}$$

subject to the condition (3), and A is again a sufficiently large bounded interval. It is assumed that nothing is known about the distribution of the responses in categories. So, uniform distributed prior information is specified for the threshold parameters, obeyed to restriction (3).

Particular parameters of the inverse-gamma distribution are selected to specify a relatively vague but proper priors for the variances of the random errors in the structural multilevel model. The random errors on different levels are assumed to be independent. The random errors on Level 2 may correlate and if prior knowledge is available it is possible to

specify this with an inverse-wishart distribution for the variance matrix $\mathbf{T}$.

Jeffreys' prior was used for the fixed effects, that is, $\gamma \sim c$, where c is a constant. The impropriety of Jeffreys' prior does not result in an improper posterior of the fixed effects.

Posterior Simulation

The likelihood in (6) involves computation of high order multidimensional integrals and makes classical inference based on maximum likelihood extremely difficult. Inference about the unknown parameters within a Bayesian framework is based on their joint posterior distribution. The joint posterior distribution of the parameters of interest is very complex but simulation based methods circumvent the computation of high dimensional integrals. An MCMC algorithm is considered to obtain random draws from the joint posterior distribution of the parameters of interest given the data. The Markov chains are relatively easy to construct and the MCMC techniques are straightforward to implement. Fox and Glas (2001, 2002) implemented a Gibbs sampler for a structural multilevel model with a latent dependent variable and a structural multilevel model with latent independent variables using dichotomous responses. The extension to a structural multilevel model with latent dependent and independent variables and dichotomous and polytomous response data is quite straightforward. The basic idea is introducing augmented data in order to draw samples from the conditional distributions of the parameters (Tanner & Wong, 1987). This has been described by Albert (1992), Albert and Chib (1993), and Johnson and Albert (1999) for the normal ogive model and the ordinal probit model, and extensively used in estimating parameters of complex models, among others, Ansari and Jedidi (2000), Béguin and Glas (2001), and Fox and Glas (2001). The full conditionals of all parameters can be specified, see Appendix A, and the Gibbs sampler is used to estimate the parameters. Each iteration of the Gibbs sampler consists of sequentially sampling from the full conditional distributions associated with the unknown parameters, $\{\omega, \xi^{\omega}, \theta, \xi^{\theta}, \beta, \sigma^2, \zeta, \xi^{\zeta}, \gamma, \mathbf{T}\}$, and sampling the augmented data to circumvent the need for integration procedures.

The convergence of the Gibbs sampling algorithm can be accelerated by using a Metropolis-Hastings step for sampling the cutoff parameters (Cowles, 1996). But constructing a suitable proposal density for the cutoff parameters can be quite difficult. Here, a new candidate is generated for cutoff parameter κ_c, the upperbound of category c, from a normal distribution,

$$\kappa_c \sim N\left(\kappa_c^{(m)}, \sigma_{MH}^2\right), \tag{9}$$

where $\kappa_c^{(m)}$ is the value of κ_c in the m^{th} iteration of the sampler. The variance of the proposal distribution, σ^2_{MH}, must be specified appropriately to establish an efficient algorithm, that is, the simulations are moving fast through the target distribution (Gelman, Roberts, & Gilks, 1996). In the present paper, the variance of this proposal distribution is adjusted within the sampling procedure. This fine tuning of the proposal distribution results in a good and efficient convergence of the algorithm without detailed prior information regarding the variance of the proposal distribution. In specific, say, after each 50^{th} iteration the acceptance rate, see Appendix A, regarding the threshold parameters is evaluated. If the acceptance rate is low, a high percentage of the sampled new candidates were rejected, the variance σ^2_{MH} is too high. The other way around, if the acceptance rate is high, a high percentage of the sampled new candidates were accepted, the variance σ^2_{MH} is too low. In both situations the variance is adjusted in the right direction. Here, the variance σ^2_{MH} is adjusted to obtain an acceptance rate of approximately .5 which was found to be optimal for univariate Metropolis-Hastings chains of certain types (Gelman, Roberts, & Gilks, 1996).

Under general conditions converges the Markov chain of sequential draws in distribution to the joint posterior distribution (Tierney, 1994). Convergence can be evaluated by comparing the between and within variance of generated multiple Markov chains from different starting points (see, for instance, Robert & Casella, 1999, pp. 366). Another method is to generate a single Markov chain and to evaluate convergence by dividing the chain into sub-chains and comparing the between- and within-sub-chain variance. A single run is less wasteful in the number of iterations needed. A unique chain and a slow rate of convergence is more likely to get closer to the stationary distribution than several shorter chains. In the examples given below, the full Gibbs sample instead of a set of sub-samples from this sample was used to estimate the parameters. The latter procedure leads to losses in efficiency (MacEachern & Berliner, 1994). Further, the CODA software (Best, Cowles, & Vines, 1995) was used to analyze the output from the Gibbs sampler and the convergence of the Markov chains. Finally, after the Gibbs sampler had reached convergence and "enough" samples were drawn, posterior means of all parameters of interest were estimated with the mixture estimator, to reduce the sampling error attributable to the Gibbs sampler (Liu, Wong, & Kong, 1994). The posterior standard deviations and highest posterior density intervals can be estimated from the sampled values obtained from the Gibbs sampler (Chen & Shao, 1999). The Appendix describes the different simulation steps and further details of the full conditional distributions.

Model Assessment

The plausibility of the model, or its general assumptions, can be assessed using posterior predictive checks (Gelman, Meng, & Stern, 1996). Let $\mathbf{y}$ be the observed data and $\mathbf{y}^{rep}$ be the replicate observations given all model parameters, denoted as $\boldsymbol{\lambda}$. Samples of the unknown model parameters are available via the MCMC algorithm. The observed data can be compared with the sampled replicated data using some test quantity or discrepancy L. The test quantity may reflect some standard checks on overall fitness or on some specific aspects of the model. A posterior predictive p-value given by

$$p(\mathbf{y}) = P\left(L(\mathbf{y}^{rep}, \boldsymbol{\lambda}) \geq L(\mathbf{y}, \boldsymbol{\lambda}) \mid \mathbf{y}, H\right) \tag{10}$$

quantifies the extremeness of an observed value of the test quantity under model H. This probability can be approximated from a sample of, say M, MCMC draws of the model parameters with

$$p(\mathbf{y}) \approx \frac{1}{M} \sum_{m=1}^{M} I\left(L\left(\mathbf{y}_{(m)}^{rep}, \boldsymbol{\lambda}_{(m)}\right) \geq L\left(\mathbf{y}, \boldsymbol{\lambda}_{(m)}\right) \mid \mathbf{y}, H\right) \tag{11}$$

where $I(.)$ denotes the indicator function. For p-values close to zero or one the posited model does not fit the data, regarding the test quantity.

An overall fit test statistics, a X^2-discrepancy as defined by Gelman, Meng, and Stern (1996), can be used to judge the fit of the model, that is,

$$L(\mathbf{y}, \boldsymbol{\lambda}) = \sum_{k=1}^{K} \sum_{i=1}^{N} \frac{(y_{ik} - E(y_{ik} \mid \boldsymbol{\lambda}))^2}{Var(y_{ik} \mid \boldsymbol{\lambda})}, \tag{12}$$

for N persons responding to K items. In fact, the X^2-discrepancy is the sum of squares of standardized residuals with respect to their expectations under the posited model. A lack of fit, a p-value close to zero or one, indicates that the observed data are not close to the replicated data under the hypothesized model H. Here, an item response theory model, as a part of the multilevel IRT model H, relates the observed data to a latent variable within the structural multilevel model. Intuitively, a lack of fit under the X^2-discrepancy provides mainly information regarding the fit of the item response theory model. In the examples below, this will turn out to be the case.

Comparing models

Bayes factors are often used when choosing between a set of competing models (see, e.g., Kass and Raftery, 1995). The underlying Bayesian argument is choosing the model for which the marginal likelihood of the data is maximal. However, there are some shortcomings regarding the Bayes factors, besides the computational problems in calculating them for high dimensional models. First, Bayes factors are not defined when using improper priors. Second, the Bayes factor tends to attach to little weight to the correct model given proper priors and an arbitrary sample size, see Gelfand and Dey (1994). Here, the pseudo-Bayes factor (PsBF) is used in comparing models which avoids these problems (Geisser & Eddy, 1979).

The pseudo-Bayes factor is based on the conditional predictive ordinate (CPO), also known as the cross-validation predictive density. Consider $i = 1, \ldots, N$ students responding to $k = 1, \ldots, K$ items. Let $\mathbf{y}_{(ik)}$ denote the observed data without a single response of student i on item k. Accordingly, the CPO is defined as

$$p\left(y_{ik} \mid \mathbf{y}_{(ik)}\right) = \int p\left(y_{ik} \mid \mathbf{y}_{(ik)}, \boldsymbol{\lambda}\right) p\left(\boldsymbol{\lambda} \mid \mathbf{y}_{(ik)}\right) d\boldsymbol{\lambda}, \tag{13}$$

$\boldsymbol{\lambda}$ represent the model parameters. It follows that $p\left(y_{ik} \mid \mathbf{y}_{(ik)}, \boldsymbol{\lambda}\right) = p\left(y_{ik} \mid \boldsymbol{\lambda}\right)$, due to conditional independence, that is, the responses on different items are independent given the ability and the responses of the students are independent of one another. These properties makes the evaluation of the cross-validation predictive density, formula (13), relatively straightforward. That is, consider $p\left(\boldsymbol{\lambda} \mid \mathbf{y}\right)$ as the importance sampling function. Given M MCMC draws of $\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(M)}$ a Monte Carlo estimate of the cross-validation predictive density (13), is given by

$$\widehat{p}\left(y_{ik} \mid \mathbf{y}_{(ik)}\right) = \left(\frac{1}{M}\sum_{m=1}^{M}\frac{1}{p^{(m)}\left(y_{ik}\right)}\right)^{-1}, \tag{14}$$

where $p^{(m)}\left(y_{ik}\right)$ is the probability on the single reponse y_{ik}, given sampled parameters $\boldsymbol{\lambda}^{(m)}$, that is, the probability for scoring correct or incorrect, formula (1) or the probability for scoring in a certain categorie on item k, formula (4). The CPO is estimated by the harmonic mean of the likelihoods using a sample from the posterior distribution $p\left(\boldsymbol{\lambda} \mid \mathbf{y}\right)$, and for $M \rightarrow \infty$ this estimate converges almost surely to the correct value (Newton & Raftery, 1994). This method can be used to estimate the pseudo-Bayes factor. The PsBF for comparing two models, H_1 and

H_2, is defined in terms of products of CPO's

$$\text{PsBF} = \prod_{i,k} \frac{p\left(y_{ik} \mid \mathbf{y}_{(ik)}, H_1\right)}{p\left(y_{ik} \mid \mathbf{y}_{(ik)}, H_2\right)}, \tag{15}$$

where y_{ik} denotes the response of student i on item k. Calculating the PsBF is straightforward using formula (14).

Most of the Bayes model assessment procedures are based on estimates of the marginal likelihood. The pseudo-Bayes factor, formula (15), is based on the observed response data. Other informal likelihood or penalized likelihood criteria can also be used for model comparison. The fit of the structural multilevel model, formula (5), can be based on the marginal likelihood of the multilevel parameters. The log-likelihood information of the multilevel parameters can be estimated using the output from the MCMC sampling scheme. An estimate of the marginal likelihood is the average of the log-likelihoods at each of the sampled points, that is,

$$\widehat{l}\left(\sigma^2, \boldsymbol{\gamma}, \mathbf{T} \mid \mathbf{y}, H\right) = \frac{1}{M} \sum_{m=1}^{M} \left(\sum_{j} \left[\sum_{i \mid j} \log p\left(\omega_{ij}^{(m)} \mid \boldsymbol{\theta}_{ij}^{(m)}, \boldsymbol{\beta}_j^{(m)}, \sigma^{2^{(m)}}, H\right) + \log p\left(\boldsymbol{\beta}_j^{(m)} \mid \boldsymbol{\zeta}_j^{(m)}, \boldsymbol{\gamma}^{(m)}, \mathbf{T}^{(m)}, H\right) \right] \right), \tag{16}$$

using the $m = 1, \ldots, M$ samples from the joint posterior distribution under model H. Instead of averaging over the log-likelihood values, another possibility could be to use the maximum log-likelihood value as an overall measure of fit, to be compared across models. In this case, the MCMC sampling run should be large to cover all possible values of the log-likelihood under the posited model.

Dempster (1997) and Aitkin (1997), considered the posterior distribution of the log-likelihood ratio (LR). The strength of evidence against model H_1 given model H_2 can be measured by (v, p_v), where p_v is the posterior probability that the LR is less than v, that is,

$$p_v = P\left(l\left(\sigma^2, \boldsymbol{\gamma}, \mathbf{T} \mid \mathbf{y}, H_1\right) - l\left(\sigma^2, \boldsymbol{\gamma}, \mathbf{T} \mid \mathbf{y}, H_2\right) < \log v \mid \mathbf{y}\right). \tag{17}$$

The case $v = 1$ is of particular importance, since $1 - p_1$ is equal to the the posterior probability that the LR is less than one. Aitkin suggests to vary v over possible values and assess changes in the posterior probability p_v that LR$< v$. The log-likelihood is a function of the data and

the parameters, and so has a posterior distribution obtainable from that of the parameters. The sampled values from the MCMC run can be used to estimate the posterior probability p_v by checking how often the inner-statement in (17) is true given the sampled log-likelihood values under both models.

Obviously, changes in the measurement model(s) and in the prior specifications are not captured by this information criterion. LR comparisons are quite insensitive to prior changes, and vary only for strongly informative priors. Below it will be shown that the LR ratio can be used to compare models with each other regarding model changes in the multilevel part. On the other hand, the PsBF, formula (15), based on the response data via an IRT model, may not always capture changes in the structural multilevel model.

Parameter Recovery

A simulation study was carried out to assess the performance of the MCMC estimation procedure. To present some empirical idea about the performance of the estimation method 100 simulated data sets were analyzed. The following structural multilevel model was considered,

$$\begin{aligned} \theta_{ij} &= \beta_{0j} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\zeta_j + u_{0j} \end{aligned} \tag{18}$$

where $e_{ij} \sim N(0, \sigma^2)$ and $u_{0j} \sim N(0, \tau^2)$. At Level 1, a sample of 2,000 students, divided equally over 200 groups, responding to a test of 40 items with four response categories was considered to measure the latent dependent variable. Responses to a test of 40 dichotomous scored items belonging to, for example, group-representatives, were considered to measure the latent Level 2 explanatory variable. For each data set, the latent abilities $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ were sampled from a standard normal distribution. The discrimination and difficulty parameters, regarding the normal ogive model for measuring $\boldsymbol{\zeta}$, were sampled as follows; $a_k \sim \log N\left(\exp(1), \frac{1}{4}\right)$ and $b_k \sim N\left(0, \frac{1}{2}\right)$, $k = 1, \ldots, 40$. The discrimination parameters in the graded response model for measuring $\boldsymbol{\theta}$ were generated according to the same distribution. The threshold parameters were choosen in such a way that the generated latent responses, according to formula (2), were divided into four response categories. The true population values of the unknown parameters, σ^2, τ^2 and $\boldsymbol{\gamma}$ are given in Table 1.

Table 1 here

For each of the 100 data sets the model parameters were estimated based on 19,000

draws from the joint posterior distribution, after a burn-in period of 1,000 draws. Initial values of the multilevel parameters were obtained by estimating the random coefficients model, formula (18), by HLM (Raudenbush, Bryk, Cheong, & Congdon, 2000) using observed sum scores as an estimate for the dependent and explanatory variable. Figure 1 shows MCMC iterates of the variance parameter at Level 1, σ^2, and variance parameter at Level 2, τ^2, of four arbitrary simulated data sets.

Figure 1 here

The left four plots correspond to the sampled values of the Level 1 variance parameter and the right four plots correspond to sampled values of the Level 2 variance parameter, for four of the simulated data sets. Visual inspection shows that the chains converged quite fast to the stationary distribution. The CODA software (Best et al., 1995) was used to check the convergence of the MCMC chains. Geweke's convergence diagnostic was computed for the several chains and p-values, given in Figure 1, indicate that the convergence of each chain is plausible. Note that the p-values were computed based on the 19,000 sampled values after the burn-in period. As an additional check, multiple chains were run from different starting points, for several simulated data sets, to verify that they resulted in similar answers. The computations were performed on a 733 MHz pentium III, written in Fortran, and each run of 20,000 iterations took about two hours.

Table 1 presents the true parameters, the average of the mean, the average of the posterior standard deviations, and the average 95% highest posterior density intervals (HPD) over the 100 MCMC samples. Further, a 95% coverage for each parameter are given in Table 1. The coverage is the proportion of the 100 HPD regions covering the true parameter values. It can be seen that there is a close agreement between the true parameters and the average estimated means, and acceptable coverage properties. Although only 100 simulated data sets were used, the average of the posterior standard deviations were comparable too the standard deviation within the 100 posterior means, for each model parameter.

Model Comparison

Two alternative models were estimated using the simulated data sets to investigate the performance of the pseudo-Bayes factor and the log-likelihood of the structural multilevel model for model comparison. The first alternative model (Model 2) corresponds to the empty model, that is, a structural multilevel model without any explanatory variables. The second alternative model (Model 3) corresponds to the model where observed sum scores were imputed for the latent dependent and explanatory variable. Accordingly, the true model will be denoted

as Model 1.

Table 2 presents the results of estimating the parameters of Model 2 and 3 using the same simulated data. Without the latent explanatory variable ζ there is a lot more unexplained variance at Level 2 but the other parameter estimates of Model 2 remain almost the same. Obviously, a higher variance at Level 2 induces a higher posterior standard deviation of the fixed effect.

Table 2 here

Model 1 and Model 2 were compared in terms of the PsBF related to the observed responses of the 2000 students on 40 items. The average PsBF across the 100 data sets for Model 1 versus Model 2 is given by $\exp(-11267 + 11268) = \exp(1)$. Although the PsBF is greater than 1, it cannot distinguish significantly Model 1 and Model 2 from each other. This follows from the fact that the estimated latent dependent variables under Model 1 and 2 are almost the same. That is, the average mean square error between the estimated latent dependent variables related to Model 1 and Model 2 over the $L = 100$ data sets is

$$\text{MSE}\left(\widehat{\boldsymbol{\theta}}_{\text{model 1}}, \widehat{\boldsymbol{\theta}}_{\text{model 2}}\right) = L^{-1}\sum_{l=1}^{L}\left[N^{-1}\sum_{i=1}^{N}\left(\widehat{\theta}_{1i}^{(l)} - \widehat{\theta}_{2i}^{(l)}\right)^2\right] \tag{19}$$

and equals .05. Here, the Level 2 explanatory variable explained variance within the latent dependent variable, but did not amount a lot of information in estimating the latent dependent variable as a parameter of the measurement model. Therefore, the pseudo-Bayes factor did not notify large differences between Model 1 and 2. The parameter estimates of the measurement model hardly changed by changing the structural multilevel model.

The difference between Model 1 and Model 2 is much better captured by the log-likelihood of the structural model. There is an explanatory variable missing in Model 2, and this had an impact on the log-likelihood of the structural model. Figure 2 displays the estimated log-likelihoods of the various models, ordered to the values of Model 1. Considering all simulated data sets, the estimated log-likelihoods of Model 1, are significantly larger than the estimated log-likehoods of Model 2, the empty model. This clearly demonstrates a preference of Model 1.

Figure 2 here

The average parameter estimates of Model 3 differ somewhat from the true parameter values. Both, the variance at Level 1 and Level 2 were too large. The scale of the latent dependent and explanatory variable in Model 1 equal the scale of the imputed observed sum scores in Model 3. As a result, the parameter estimates are comparable and the same amount

of variance can be explaind by both models. The observed sum scores displayed less variance between students than the students' item responses. Accordingly, the covariate at Level 2 explains less variance between groups, and its coefficient is under-estimated. The estimates of the variance at Level 1 and Level 2 are somewhat higher but the same amount of variance is available in the dependent variable. So, Model 1 explains more variance and fits the data better. Although the differences between log-likelihoods are small, in Figure 2, it can be seen that overall Model 1 performs better than Model 3. The posterior probability of the LR ratio of Model 3 against Model 1, formula (17), was estimated, and the mean across the 100 datasets equals for $v = .1$ and $v = 1$, $p_{.1} = .150$ and $p_1 = .210$, respectively. This provides evidence that the LR is larger than one, indicating that Model 1 should be preferred above Model 3. The mean square error, as defined in (19), between the true simulated abilities, $\boldsymbol{\theta}$, and $\widehat{\boldsymbol{\theta}}_{\text{model 1}}$ equals .04, whereas the mean square error between the true simulated abilities and $\widehat{\boldsymbol{\theta}}_{\text{model 3}}$ equals .62. The simulated distributions of the latent variables, $\boldsymbol{\theta}$, ζ were both normal distributed. In Fox and Glas (2002) it was shown that the differences between the observed sum scores and the estimated abilities using IRT were much larger for skewed latent distributions.

Analyzing Multilevel Data with Measurement Error

The multilevel IRT model was used in the analysis of a mathematics test, administered to 3,500 grade 7 students in 119 schools located in the West Bank. The mathematics test consisted of 50 dichotomous scored items. Main interest was focused on exploring differences within and between schools in the West Bank and establishing factors which explain these differences with respect to students' mathematic abilities. Therefore, various background variables were measured. That is, characteristics of students, teachers, and schools were administered. Besides the mathematics and language test, an intelligence test (IQ) was administered, gender was recorded, as zero for male and one for woman, and socio-economic status (SES) was measured by the educational level of the parents. In the analyses, the observed sum scores of the predictors IQ and SES were standardized.

Tests were taken by teachers and school principles to measure aspects, as the school climate and leadership of the principle. The school climate (Climate), from teacher's perspective, was measured by 23 5-point Likert items, and leadership (Leader) was measured by 25 5-point Likert items. In the sampling design only one class was selected from each school, so the data comprehended a student (Level 1) and school level (Level 2). A stratified sample of schools ensured that all school types and all geographical districts were represented.

The average number of students per class is 28, with a minimum of 10 and a maximum of 46 students. A complete description of the data, including the data collection procedure and the different questionnaires, can be found in Shalabi (2002).

The variation in the test results of the mathematic items were modeled in terms of single underlying abilities. That is, a two-parameter normal ogive model was used to define the relationship between the observed responses and the latent dependent abilities in the stuctural multilevel model. First, the variation in the math-abilities and heterogeneity across schools was measured with an empty stuctural multilevel model, that is, only an intercept at Level 1 varying across schools. Second, student characteristics were used as predictors to explain variation. Third, the latent school characteristics, school climate and leadership, were used as Level 2 predictors on the Level 1 intercept.

The developed MCMC estimation procedure was applied to estimate the parameters of the various models. All models were identified by transforming the scale of the latent variables to a standardized normal scale. This way, the estimated parameters and log-likelihoods were comparable. The convergence of the MCMC chains was monitored by comparing the between and within variance of the generated Markov chains. Further, Geweke's convergence diagnostic was computed for the several chains and indicated that chains of 50,000 iterations had converged after a burn-in period of 1,000 iterations.

The empty model is called Model 1, and the structural multilevel model including the three Level 1 predictors is called Model 2. Model 2 is given by,

$$\begin{aligned}
\theta_{ij} &= \beta_{0j} + \beta_{1j}\text{SES}_{ij} + \beta_{2j}\text{Gender}_{ij} + \beta_{3j}\text{IQ}_{ij} + e_{ij} \qquad (20)\\
\beta_{0j} &= \gamma_{00} + u_{0j}\\
\beta_{1j} &= \gamma_{10}\\
\beta_{2j} &= \gamma_{20}\\
\beta_{3j} &= \gamma_{30}
\end{aligned}$$

where the error terms e_{ij} and u_{0j} are independent and normally distributed with zero mean and variances σ^2 and τ^2, respectively. The two-parameter normal ogive model was used to measure the latent dependent variable. The parameter estimates of Model 1 and 2 are given in Table 3.

Table 3 here

Due to scaling is the population mean or grand mean of the math abilities, γ_{00}, zero. The estimated intra-school correlation coefficient, from Model 1, is around .50, which means

that around 50% of the total variance, due to individual differences in math abilities, can be explained by school differences. For example, the difference between the nine worst and eight best performing schools is 30% items correct. The three Level 1 predictors all have a positive significant effect on students' mathematics achievement. The math abilities were scaled around zero, so, female students performed better than male students. From Table 3 it can be seen that the three Level 1 variables together account for a substantial proportion of variation in students' achievement: 21% of the student level and 27% of the school level variance.

The relevance of the three Level 1 predictors is supported by the pseudo-Bayes factor and the log-likelihood values of both models. The estimated pseudo-Bayes factor in favor of Model 2 is $\exp(-96773 + 96837) = \exp(64)$ and provide strong evidence that Model 2 fits the data better. Besides, the log-likelihood of the structural multilevel model went up from -7285.9 to -6383.8. The p-value of the overall fit test statistic, formula (10), related to the observed item responses, was around .5 for both models.

Model 2 was extended by including two latent predictors at Level 2, Leadership and Climate. The estimated multilevel IRT model (Model 3) consists of three measurement models, a two-parameter normal ogive model for measuring the latent depedendent variable, and two graded response models for measuring the latent variables at Level 2 using the polytomous scored item reponses. The structural multilevel part is given by

$$\begin{aligned}
\theta_{ij} &= \beta_{0j} + \beta_{1j}\text{SES}_{ij} + \beta_{2j}\text{Gender}_{ij} + \beta_{3j}\text{IQ}_{ij} + e_{ij} \qquad (21)\\
\beta_{0j} &= \gamma_{00} + \gamma_{01}\text{Climate}_j + \gamma_{02}\text{Leader}_j + u_{0j}\\
\beta_{1j} &= \gamma_{10}\\
\beta_{2j} &= \gamma_{20}\\
\beta_{3j} &= \gamma_{30}
\end{aligned}$$

where the explanatory variables at Level 2, Climate and Leader, are latent explanatory variables, providing information regarding the management characteristics of the schools. All parameters were estimated simultaneously using the developed MCMC sampler. The estimated multilevel parameters are given in Table 4.

Table 4 here

Both predictors at Level 2 are significant, with a 5% significance level. The variable Leader has a positive effect on the math abilities, but the effect of variable Climate is negative. Both variables account for 8% of the school level variance. The p-values of the X^2-discrepancy,

corresponding to the Level 2 variables, were around .8, meaning that the averaged sum over the standardized residuals based on predictive data were somewhat higher than the sum over the standardized residuals based on the observed data. That is, the estimated graded response models did not replicate data close to the observed data. The difference in the log-likelihood, from -6383.8 (Model 2) to -6246.8, is not significant given the 95% highest posterior density interval $[-6452.1, -6034.6]$ for the log-likelihood of Model 3. The posterior probability p_v, defined in (17), for the LR of Model 3 against Model 2 equals .865 for $v = .1$. This means that the posterior probability that the LR is less than .1 equals .865. Also, the pseudo-Bayes factor, related to the observed data for measuring the latent dependent variable, did not show a preference for Model 3. It turns out that the latent variables at Level 2, with significant coefficients, did not result in a better model fit. The school organisational and instructional variables, school climate and school leadership, are rarely investigated in developing countries and proved to have not much of an influence on the students' mathematics abilities.

Analogous to a standard multilevel analysis, observed sum scores were used as estimates for the latent mathematic abilities and the latent school variables, Climate and Leader. Then, the Gibbs sampler, described in Appendix A, was used to estimate the parameters of Model 3, formula (21), where the observed sum scores were scaled the same way as the latent variables within the multilevel IRT Model 3. The parameter estimates are shown in Table 4. It can be seen that the parameter estimates are lower than the estimates resulting from the multilevel IRT model analysis, due to measurement error in the observed sum scores. The estimate of the variance at Level 1 is higher and at Level 2 is lower, meaning that there is more unexplained variance using observed sum scores. Less variance is explained due to differences between schools and less variance is explained by the Level 1 characteristics, SES, gender and IQ. The latent dependent variable measured with an IRT model displays more differences between students than the observed sum scores, that is, the observed sum scores display less variance between students than the students' item responses. The effects of the Level 2 variables were lower when observed sum scores were used, both effects are still significant. As in the corresponding multilevel IRT analysis, the estimated log-likelihood of Model 3 is not significantly higher than the estimated log-likelihood of Model 2 using observed sum scores, from -6445.3 (Model 2) to -6368.6. The estimates are smaller than the corresponding multilevel IRT log-likelihoods. The log-likelihood of the structural multilevel model is maximized using the multilevel IRT model, in spite of a poor fit of the graded response models.

In sum, primary schools in the West Bank differ a lot, considering the math abilities of

the students, but the school context, measured by climate and leader, did not explain much variation at Level 2. The Level 1 characteristics SES, IQ and gender explained a lot of variation at the student level. There was an increase in the effects of school characteristics on students' achievements in comparison to traditional methods for analyzing these data. Modeling measurement error in the latent dependent and independent explanatory variables resulted in larger effects and more explained variance at both levels. The effects were attenuated when traditional methods were used which ignored the measurement error, that is, using observed sum scores as an estimate for the latent variables.

Conclusions

A multilevel IRT model has been proposed that contains latent dependent and/or explanatory variables on different levels. Item response theory models are used to define the relationship between observable test scores and the latent constructs. The model can handle dichotomous and polytomous responses. The structural multilevel model describes the relationship between different latent constructs and observed variables on different levels.

The simulation study shows that the Bayesian estimation method works well. The MCMC algorithm is very flexible and allows the modeling of various latent variables on different levels using dichotomous and/or polytomous responses. The flexibility of the estimation procedure allows the use of other measurement error models and can handle multilevel models with three or more levels. The estimation procedure takes the full error structure into account and allow for errors in both the dependent and independent variables. The Metropolis-Hastings algorithm is used to sample parameters via a proposal distribution from which it is easy to sample. A good convergence of the algorithm is obtained by adjusting the variance of the proposal distribution. The developed Bayesian estimation method for estimating all parameters simultaneously is implemented in Fortran and freely available (Fox , 2003). The program runs within the statistical package S-plus (Insightful, 2001).

The posterior predicticve checking provides information regarding the global fit of the model. Within the framework of the posterior predictive checks, other specific diagnostics can be developed to check assumptions as local independence, heteroscedasticity, and autocorrelation. Since the MCMC run can be time-consuming, it contains the estimation of the model parameters and the checking of some of the model assumptions. Various applications and developments of complex psycometric models show this twofold use of the MCMC samples, see, for example, Ansari and Jedidi (2000), Béguin and Glas (2001), and, Lee and

Zhu (2000).

The pseudo-Bayes factor can be used to compare models with each other but it is sensitive to the prior choice, and may not always reflect changes within the structural model. Therefore, modeling differences within the structural part are better assessed by looking at the likelihood of the structural part. The complex likelihood of the multilevel IRT model reveals the usefulness of looking at a part of the likelihood. The log-likelihood quantity could be extended to penalise models which improve fit at the expense of more parameters, and so serves as a measure to assess model parsimony. For example, a Bayesian Information Criterium (BIC) could be defined to compare multilevel IRT models with different structural multilevel parts.

It is hard to give a general specification of when the multilevel IRT model will make a substantive difference in the analysis, besides the theoretical considerations. In cases of skewed distributions or cases where some of the responses to the items are missing the multilevel IRT model is preferred. In case of missing response data, the MCMC estimation procedure for complete data can be modified in such a way that only the available data are used. This is done by defining an indicator variable that specifies the items that are administered, and the persons who are responding. The example showed a better fit of the multilevel IRT model. In case of a smaller number of Level 1 units or response items, or bad fit of one of the measurement models, a multilevel model with observed sum scores could be preferred. In general, more research is needed to obtain rules for choosing between these models in different situations.

In the present paper, the measurement models, within the multilevel IRT model, assume that the ability parameter is unidimensional. In some situations, a priori information may show that multiple abilities are involved in producing the observed response patterns. Then, a multidimensional IRT model serves to link the observed response data to several latent variables. The multilevel IRT model could be extended to handle these correlated latent variables within the structural multilevel model. Two options are possibile, one of the correlated latent variables is a dependent variable or all latent variables are explanatory variables within the structural multilevel model. This way, the dependency structure and other person and group characteristics can be taken into account in analysing the relation between multidimensional latent abilities. The parameters of a normal ogive multidimensional IRT models can be estimated within a Bayesian framework using the Gibbs sampler (Béguin & Glas, 2001). Accordingly, the parameters of this extended multilevel IRT model can be estimated within a Bayesian framework using MCMC, by defining the full conditionals of all parameters.

Appendix A. The MCMC Implementation

The Gibbs sampler consists of stepwise draws from the full conditional distributions. The algorithm is specified by defining all the full conditional distributions. Accordingly, the $(m+1)^{th}$ iteration involves generating draws from these distributions. Below, an implementation is given for an arbitrary latent variable in the structural multilevel model. In all the steps, other possible latent variables are treated as observed variables. Obviously, the full conditionals of other latent variables and parameters of the corresponding measurement models can be obtained in the same way.

The first step is to augment the observed data, $\mathbf{y}$, with latent data $\mathbf{z}$. By defining a continuous latent variable, z, that underlies the binary or polytomous response it is easier to sample from the conditional distributions of the parameters of interest. This augmented data, as defined in formula (2) and below formula (1), serve to simplify calculations. This procedure has been widely applied, see, for example, Albert (1992), and Johnson and Albert (1999). Let $\mathbf{z}$ denote the augmented data regarding the observed binary or polytomous data, $\mathbf{y}$, for measuring the latent ability $\boldsymbol{\theta}$. Accordingly, let $\boldsymbol{\theta}$ be an arbitrary latent variable within the structural multilevel model.

(1) The conditional distribution of the discrimination and difficulty parameters in the normal ogive model, formula (1), can be obtained by viewing these parameters as coefficients in the regression of $\mathbf{z}$ on $\mathbf{H} = [\boldsymbol{\theta}, -\mathbf{1}]$. It follows that,

$$\boldsymbol{\xi}_k \mid \boldsymbol{\theta}, \mathbf{z}_k \sim N\left(\widehat{\boldsymbol{\xi}}_k, \left(\mathbf{H}^t\mathbf{H}\right)^{-1}\right) I\left(a_k > 0\right) I\left(a_k \in A\right), \tag{22}$$

where $\boldsymbol{\xi}_k = (a_k, b_k)$, and A a sufficiently large bounded interval. The full conditional distribution of the discrimination parameter in the graded response model, formula (4), can be obtained in the same way.

(2) The conditional distribution of the threshold parameter is difficult to specify. Therefore, a candidate κ_k^*, regarding the thresholds of item k, is sampled from a proposal distribution, formula (9), from which it is easy to sample. The candidate is accepted or rejected based

on the Metropolis-Hastings acceptance probability,

$$\min\left[\prod_{ij}\frac{\Phi\left(\kappa^*_{ky_{ij,k}}-a_k\theta_{ij}\right)-\Phi\left(\kappa^*_{ky_{ij,k}-1}-a_k\theta_{ij}\right)}{\Phi\left(\kappa_{ky_{ij,k}}-a_k\theta_{ij}\right)-\Phi\left(\kappa_{ky_{ij,k}-1}-a_k\theta_{ij}\right)}\prod_{c=1}^{C_k-1}\frac{\Phi\left(\kappa_{kc+1}-\kappa_{kc}\right)/\sigma_{MH}-\Phi\left(\kappa^*_{kc-1}-\kappa_{kc}\right)/\sigma_{MH}}{\Phi\left(\kappa^*_{kc+1}-\kappa^*_{kc}\right)/\sigma_{MH}-\Phi\left(\kappa_{kc-1}-\kappa^*_{kc}\right)/\sigma_{MH}},1\right]$$

where $y_{ij,k}$ denotes the response of person ij on item k. For the other parameters the sampled values from the last iteration are used. The first part represents the contribution from the likelihood whereas the second part represents normalized proposal distributions.

(3) The conditional distribution of the latent variable $\boldsymbol{\theta}$. The latent variable is a dependent variable or an independent variable at Level 1 or Level 2 in the structural multilevel model. In all three cases, the conditional distribution is a product of two normal distributions and the full conditional distribution follows from standard properties of normal distributions (Lindley & Smith, 1972). In all cases, one part follows from the measurement model, where θ_{ij} can be viewed as a regression coefficient in the regression from $z_{ijk}-b_k$ or z_{ijk} on a_k in case of a binary or polytomous data, respectively. Here, the three separate cases are described using the graded response model.

- Dependent latent variable θ_{ij}. It follows from, formula (2) and (5), that

$$\theta_{ij}\mid \mathbf{z}_{ij},\boldsymbol{\xi},\boldsymbol{\beta}_j,\sigma^2,\mathbf{y}\sim N\left(\frac{\widehat{\theta}_{ij}/v+\mathbf{X}_{ij}\boldsymbol{\beta}_j/\sigma^2}{1/v+1/\sigma^2},\frac{1}{1/v+1/\sigma^2}\right), \tag{23}$$

  with $\widehat{\theta}_{ij}=\sum_k a_k z_{ijk}/\sum_k a_k^2$ and $v=1/\sum_k a_k^2$.
- Explanatory latent variable θ_{ij} at Level 1. Again, from formula (2) and (5), it follows that

$$\theta_{ij}\mid \mathbf{z}_{ij},\boldsymbol{\xi},\boldsymbol{\beta}_j,\sigma^2,\mathbf{y}\sim N\left(\frac{\widehat{\theta}_{ij}/v+\widetilde{\theta}_{ij}/\phi}{1/v+1/\phi},\frac{1}{1/v+1/\phi}\right), \tag{24}$$

  where the posterior expectation constitutes of $\widehat{\theta}_{ij}$, as defined above, and a term $\widetilde{\theta}_{ij}=\beta_{qj}^{-1}\left(\omega_{ij}-\boldsymbol{\beta}_j^{-}\mathbf{X}_{ij}^{-}\right)$, and the posterior variances of v and $\phi=\beta_{qj}^{-2}\sigma^2$, where β_{qj} is the regression coefficent of θ_{ij} and $\boldsymbol{\beta}_j^{-}\mathbf{X}_{ij}^{-}$ the product of regression coefficients and explanatory variables at Level 1 witout the latent variable θ_{ij}.

– Explanatory latent variable θ_j at Level 2. In the same way, it follows that

$$\theta_j \mid \mathbf{z}_j, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_q, \mathbf{T}, \mathbf{y} \sim N\left(\frac{\widehat{\theta}_j/v + \widetilde{\theta}_j/\phi}{1/v + 1/\phi}, \frac{1}{1/v + 1/\phi}\right), \tag{25}$$

where again $\widehat{\theta}_j$ is the least squares estimator following from the measurement model, formula (2), and $\widetilde{\theta}_j = \gamma_{qs}^{-1}\left(\beta_{qj} - \boldsymbol{\gamma}_q^{-}\mathbf{W}_j^{-}\right)$ with $\phi = \mathbf{T}_{qq}/\gamma_{qs}^2$, where γ_{qs} is the regression coefficient of explanatory variable θ_j, and $\boldsymbol{\gamma}_q^{-}\mathbf{W}_j^{-}$ is the product of other regression coefficients and explanatory variables. When defining a normal distributed prior for $\boldsymbol{\theta}$, formulae (23), (24), and (25) are easily extended, see Fox and Glas (2002).

(4) The full conditional for the regression coefficient, $\boldsymbol{\beta}_j$. Let $\mathbf{X}$ and $\mathbf{W}$ be the explanatory variables at Level 1 and 2, respectively, including any latent explanatory variables. From formula (5), and a noninformative prior, it follows that

$$\boldsymbol{\beta}_j \mid \sigma^2, \boldsymbol{\gamma}, \mathbf{T}, \mathbf{y} \sim N\left(\frac{\mathbf{X}_j^t\mathbf{X}_j\widehat{\boldsymbol{\beta}}_j/\sigma^2 + \mathbf{T}^{-1}\mathbf{W}_j\boldsymbol{\gamma}}{\mathbf{X}_j^t\mathbf{X}_j/\sigma^2 + \mathbf{T}^{-1}}, \frac{1}{\mathbf{X}_j^t\mathbf{X}_j/\sigma^2 + \mathbf{T}^{-1}}\right), \tag{26}$$

where $\widehat{\boldsymbol{\beta}}_j = \left(\mathbf{X}_j^t\mathbf{X}_j\right)^{-1}\mathbf{X}_j^t\boldsymbol{\omega}_j$. The matrix $\mathbf{X}_j$ does not need to be of full rank, since the inverse of $\mathbf{X}_j^t\mathbf{X}_j$ is not needed.

(5) The full conditional for the fixed effects, $\boldsymbol{\gamma}$. Again, $\mathbf{W}$ represents the explanatory variables at Level 2, including the latent variables at Level 2. From formula (5), and a noninformative prior, it follows that

$$\boldsymbol{\gamma} \mid \boldsymbol{\beta}_j, \mathbf{T}, \mathbf{y} \sim N\left(\frac{\sum_j \mathbf{W}_j^t\mathbf{T}^{-1}\boldsymbol{\beta}_j}{\sum_j \mathbf{W}_j^t\mathbf{T}^{-1}\mathbf{W}_j}, \frac{\mathbf{1}}{\sum_j \mathbf{W}_j^t\mathbf{T}^{-1}\mathbf{W}_j}\right). \tag{27}$$

(6) The full conditional for the variance at Level 1, σ^2. A prior for the variance can be specified in the form of an inverse-gamma (IG) distribution with shape and scale parameters, $(n_0/2, n_0S_0/2)$. S_0 is a prior quess and n_0 displayes the strength of this belief. It follows that

$$\sigma^2 \mid \boldsymbol{\beta}, \mathbf{y} \sim \mathrm{IG}\left(\frac{N + n_0}{2}, \frac{NS + n_0S_0}{2}\right), \tag{28}$$

where $S = \sum_{i|j} 1/n_j \left(\omega_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_j\right)^2$. A non-informative but proper prior is specified if $n_0 = .0001$ and $S_0 = 1$ (Congdon, 2002).

(7) The full conditional for the variance at Level 2, $\mathbf{T}$. An inverse Wishart distribution with small degrees of freedom, but greater than the dimension of $\boldsymbol{\beta}_j$, n_0, and unity-matrix, $\mathbf{S}_0$,

can be used as a diffuse proper prior for $\mathbf{T}$. It follows that

$$\mathbf{T} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{y} \sim \text{Inv-Wishart}\left(n_0 + J, (\mathbf{S} + \mathbf{S}_0)^{-1}\right) \tag{29}$$

where $\mathbf{S} = \sum_j (\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma})(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma})^t$.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variable regression. *Journal of Educational and Behavioral Statistics, 22*, 47-76.

Aitkin, M. (1997). The calibration of p-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood. *Statistics and Computing, 7*, 253-261.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17*, 251-269.

Albert, J.H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88*, 669-679.

Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika, 65*, 475-496.

Béguin, A. A., & Glas, C.A.W. (2001). MCMC estimation of multidimensional IRT models. *Psychometrika, 66*, 541-562.

Best, N. G., Cowles, M. K., & Vines, S. K. (1995). CODA Convergence diagnosis and output analysis software for Gibbs Sampler output: Version 0.3 [Computer software and manual]. Cambridge, UK: Biostatistics Unit-MRC.

Bock, R.D. (Ed.) (1989). *Multilevel analysis of educational data.* San Diego, CA: Academic Press, Inc.

Chen, M. -H., & Shao, Q. -M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics, 8*, 69-92.

Congdon, P. (2002). *Bayesian Statistical Modeling.* New York, NY: John Wiley & Sons, Inc.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing, 6*, 101-111.

Dempster, A. P. (1997). The direct use of likelihood for significance testing. *Statistics and Computing, 7*, 247-252.

Fox, J. -P. (2003). *Multilevel IRT Manual.* Technical Report (in press). Faculty of Educational Science and Technology, Twente University, Enschede. Downloadable from http://users.edte.utwente.nl/Fox.

Fox, J. -P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 269-286.

Fox, J. -P., & Glas, C. A. W. (2002). Bayesian modeling of measurement error in predictor variables using Item Response Theory. *Psychometrika,* in press.

Geisser, S., & Eddy, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association, 74*, 153-160.

Gelfand, A.E., & Dey, D. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B, 56*, 501-514.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association, 85*, 972-985.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Gelman, A., Meng X. -L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733-807.

Gelman A., Roberts, G. O., & Gilks, W. R. (1996). Efficient Metropolis Jumping Rules. In J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith (Eds.), *Bayesian Statistics 5* (pp. *599-607).* Oxford University Press.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold.

Insightful Corporation (2001). *S-Plus 6 for Windows.* Seattle, WA.

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling.* New York, NY: Springer-Verlag, Inc.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79-93.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

Lee, S.-Y., Poon, W.-Y., & Bentler, P.M. (1992). Structural equation models with

continuous and polytomous variables. *Psychometrika, 57*, 89-106.

Lee, S.-Y., & Zhu, H.-T. (2000). Statistical analysis of non-linear equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychologicaly, 53*, 209-232.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B, 34*, 1-41.

Liu, J. S., Wong, H. W., & Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika, 81*, 27-40.

MacEachern, S. N., & Berliner, L. M. (1994). Subsampling the Gibbs sampler. *The American Statistician, 48*, 188-190.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics, 26*, 307-330.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19*, 91-100.

Newton, M. & Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B, 56*, 3-48.

Muraki, E. & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*, 73-90.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T., Jr. (2000). *HLM 5. Hierarchical linear and nonlinear modeling.* Lincolnwood, IL; Scientific Software International, Inc.

Raudenbush, S. W., & Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology, 29*, 1-41.

Robert, C. P., & Casella, G. (1999). *Monte Carlo statistical methods.* New York, NY: Springer.

Shalabi, F. (2002). *Effective schooling in the West Bank.* Unpublished doctoral dissertation, Twente University, Enschede, Netherlands.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*, 528-550.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of*

Statistics, 22, 1701-1762.

Samejima, F. (1969). Estimation of a Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika Monograph Supplement, No. 17.*

Verhelst, N. D., & Eggen, T. J. H. M. (1989). *Psychometrische en Statistische Aspecten van Peilingsonderzoek* (PPON rapport 4) (In Dutch). Arnhem: Cito.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika, 56*, 589-600.

Zwinderman, A. H. (1997). Response models with manifest predictors. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern Item Response Theory* (pp. 245-256). New York, NY: Springer.

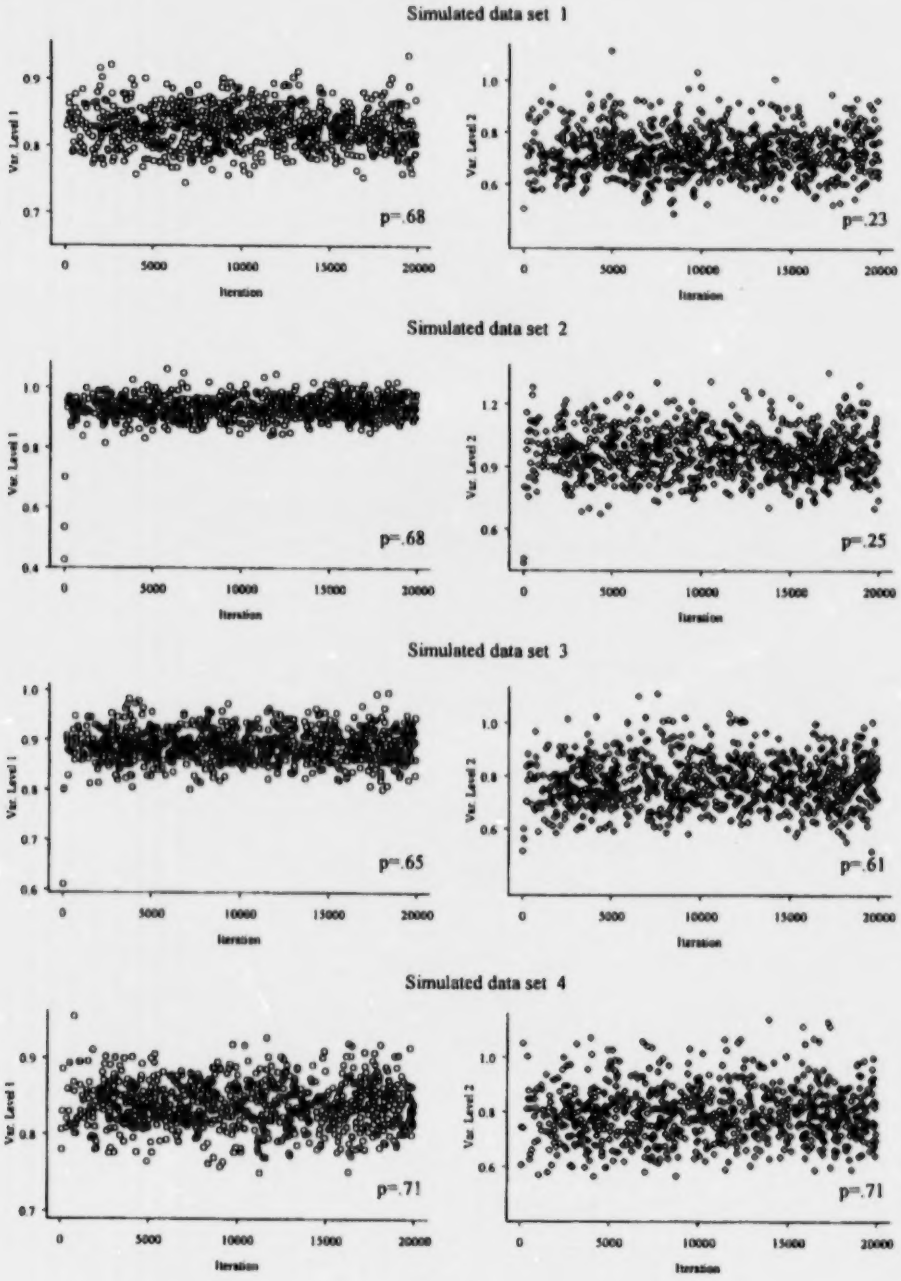


Figure 1. MCMC iterations of the variance parameters corresponding to the multilevel IRT model. The p-values correspond to the Geweke convergence diagnostic.

BEST COPY AVAILABLE

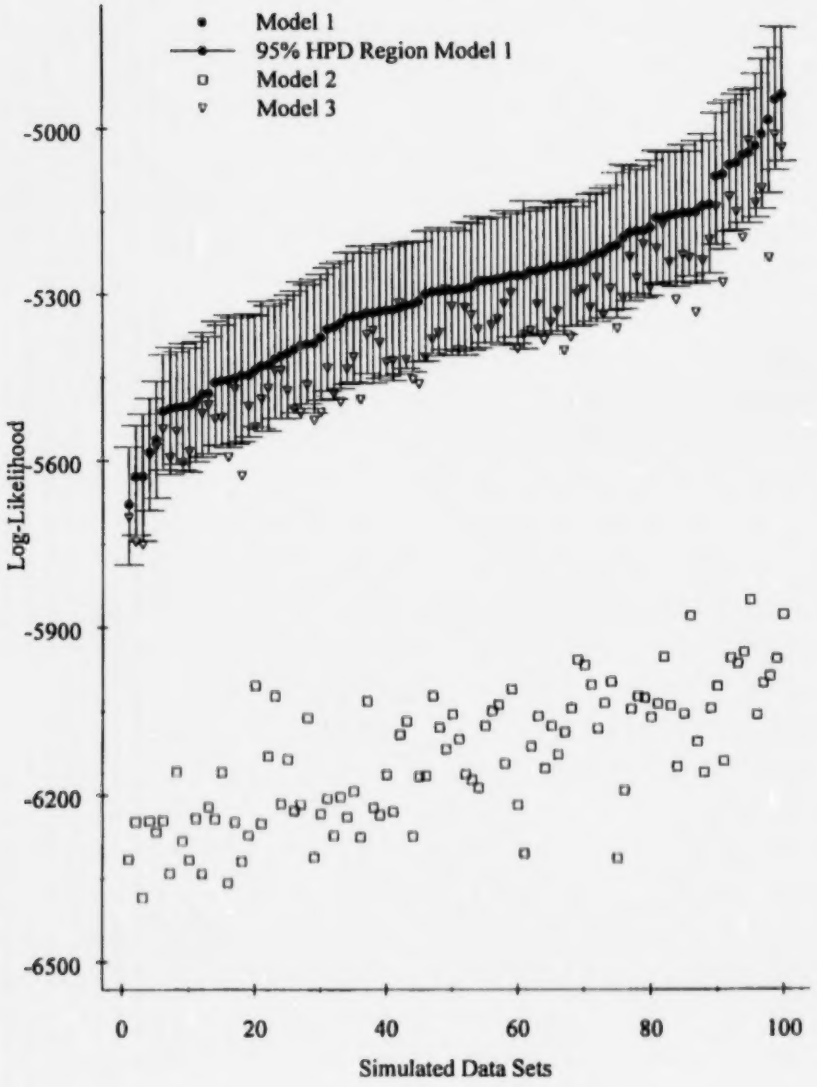


Figure 2. The estimated log-likelihoods of the structural multilevel part of Model 1, Model 2 and Model 3.

Table 1. Generating values, means and standard errors of recovered values

	Generated	Multilevel IRT			
Fixed Effects	Coeff.	Mean of Estimates	Standard Deviation	HPD	Coverage
γ_{00}	1.25	1.254	.069	[1.122, 1.387]	.94
γ_{01}	1	.996	.069	[.858, 1.127]	.96
Random Effects	Var. Comp.	Var. Comp.	Standard Deviation	HPD	Coverage
σ^2	.9	.900	.033	[.837, .964]	.86
τ^2	.75	.780	.095	[.600, .967]	.92

Table 2. Parameter estimates of two alternative models

	Emty Model				Multilevel Model			
Fixed Effects	Mean of Estimates	Standard Deviation	HPD	Covr.	Mean of Estimates	Standard Deviation	HPD	Covr.
γ_{00}	1.249	.108	[1.053, 1.457]	.99	1.247	.067	[1.114, 1.377]	.96
γ_{01}	–	–	–	–	.934	.067	[0.780, 1.063]	.84
Random Effects	Variance Components	Standard Deviation	HPD	Covr.	Variance Components	Standard Deviation	HPD	Covr.
σ^2	.902	.033	[.838, .966]	.86	.940	.031	[0.878, 1.001]	.67
τ^2	1.780	.191	[1.424, 2.163]	0	.809	.092	[0.637, .991]	.93

Table 3. Parameter estimates of Model 1 and 2

	Model 1			Model 2		
Fixed Effects	Coefficient	Standard Deviation	HPD	Coefficient	Standard Deviation	HPD
γ_{00}	.005	.066	[−.130, .131]	−.097	.064	[−.224, .028]
γ_{10} (SES)	–	–	–	.124	.015	[.096, .153]
γ_{20} (Gender)	–	–	–	.213	.061	[.093, .333]
γ_{30} (IQ)	–	–	–	.351	.015	[.322, .380]
Random Effects	Variance Components	Standard Deviation	HPD	Variance Components	Standard Deviation	HPD
σ^2	.515	.014	[.487, .543]	.408	.012	[.385, .432]
τ^2	.507	.069	[.378, .646]	.370	.052	[.282, .484]

Table 4. Parameter estimates of multilevel IRT Model 3, and multilevel Model 3 using observed sum scores

	Model 3			Model 3 (sum scores)		
Fixed Effects	Coefficient	Standard Deviation	HPD	Coefficient	Standard Deviation	HPD
γ_{00}	−.095	.063	[−.218, .027]	−.088	.059	[−.205, .026]
γ_{01} (Climate)	.238	.084	[.070, .401]	.205	.075	[.056, .350]
γ_{02} (Leader)	−.126	.085	[−.298, .035]	−.119	.075	[−.266, .028]
γ_{10} (SES)	.125	.015	[.096, .153]	.111	.014	[.084, .139]
γ_{20} (Gender)	.211	.061	[.095, .332]	.193	.055	[.084, .299]
γ_{30} (IQ)	.351	.015	[.322, .381]	.341	.015	[.311, .368]
Random Effects	Variance Components	Standard Deviation	HPD	Variance Components	Standard Deviation	HPD
σ^2	.408	.012	[.385, .432]	.471	.012	[.448, .494]
τ^2	.340	.050	[.248, .438]	.314	.044	[.235, .405]

RR-01-04 R. Ben-Yashar, S. Nitzan & H.J. Vos, *Optimal Cutoff Points in Single and Multiple Tests for Psychological and Educational Decision Making*

RR-01-03 R.R. Meijer, *Outlier Detection in High-Stakes Certification Testing*

RR-01-02 R.R. Meijer, *Diagnosing Item Score Patterns using IRT Based Person-Fit Statistics*

RR-01-01 H. Chang & W.J. van der Linden, *Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach*

RR-00-11 B.P. Veldkamp & W.J. van der Linden, *Multidimensional Adaptive Testing with Constraints on Test Content*

RR-00-10 W.J. van der Linden, *A Test-Theoretic Approach to Observed-Score Equating*

RR-00-09 W.J. van der Linden & E.M.L.A. van Krimpen-Stoop, *Using Response Times to Detect Aberrant Responses in Computerized Adaptive Testing*

RR-00-08 L. Chang & W.J. van der Linden & H.J. Vos, *A New Test-Centered Standard-Setting Method Based on Interdependent Evaluation of Item Alternatives*

RR-00-07 W.J. van der linden, *Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing*

RR-00-06 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory*

RR-00-05 B.P. Veldkamp, *Modifications of the Branch-and-Bound Algorithm for Application in Constrained Adaptive Testing*

RR-00-04 B.P. Veldkamp, *Constrained Multidimensional Test Assembly*

RR-00-03 J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*

RR-00-02 J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*

RR-00-01 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.

BEST COPY AVAILABLE

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-02-11 J.P. Fox, *Multilevel IRT Using Dichotomous and Polytomous Response Data*

RR-02-10 H.J. Vos, *Applying the Minimax Principle to Sequential Mastery Testing*

RR-02-09 B.P. Veldkamp, W.J. van der Linden & A. Ariel, *Mathematical-Programming Approaches to Test Item Pool Design*

RR-02-08 B.P. Veldkamp, *Optimal Test Construction*

RR-02-07 B.P. Veldkamp & A. Ariel, *Extended Shadow Test Approach for Constrained Adaptive Testing*

RR-02-06 W.J. van der Linden & B.P. Veldkamp, *Constraining Item Exposure in Computerized Adaptive Testing with Shadow Tests*

RR-02-05 A. Ariel, B.P. Veldkamp & W.J. van der Linden, *Constructing Rotating Item Pools for Constrained Adaptive Testing*

RR-02-04 W.J. van der Linden & L.S. Sotaridona, *A Statistical Test for Detecting Answer Copying on Multiple-Choice Tests*

RR-02-03 W.J. van der Linden, *Estimating Equating Error in Observed-Score Equating*

RR-02-02 W.J. van der Linden, *Some Alternatives to Sympson-Hetter Item-Exposure Control in Computerized Adaptive Testing*

RR-02-01 W.J. van der Linden, H.J. Vos, & L. Chang, *Detecting Intrajudge Inconsistency in Standard Setting using Test Items with a Selected-Response Format*

RR-01-11 C.A.W. Glas & W.J. van der Linden, *Modeling Variability in Item Parameters in Item Response Models*

RR-01-10 C.A.W. Glas & W.J. van der Linden, *Computerized Adaptive Testing with Item Clones*

RR-01-09 C.A.W. Glas & R.R. Meijer, *A Bayesian Approach to Person Fit Analysis in Item Response Theory Models*

RR-01-08 W.J. van der Linden, *Computerized Test Construction*

RR-01-07 R.R. Meijer & L.S. Sotaridona, *Two New Statistics to Detect Answer Copying*

RR-01-06 R.R. Meijer & L.S. Sotaridona, *Statistical Properties of the K-index for Detecting Answer Copying*

RR-01-05 C.A.W. Glas, I. Hendrawan & R.R. Meijer, *The Effect of Person Misfit on Classification Decisions*

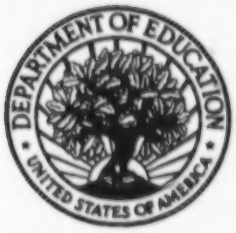
BEST COPY AVAILABLE

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

BEST COPY AVAILABLE

TM034732



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

NOTICE

Reproduction Basis

This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").